SCIENCE CHINA
Life Sciences

• RESEARCH PAPER •

August 2013 Vol.56 No.8: 751–757
doi: 10.1007/s11427-013-4500-6

Predicting potential cancer genes by integrating network properties, sequence features and functional annotations

LIU Wei & XIE HongWei*

College of Mechanical & Electronic Engineering and Automatization, National University of Defense Technology, Changsha 410073, China

Received October 25, 2012; accepted May 14, 2013; published online July 4, 2013

The discovery of novel cancer genes is one of the main goals in cancer research. Bioinformatics methods can be used to accelerate cancer gene discovery, which may help in the understanding of cancer and the development of drug targets. In this paper, we describe a classifier to predict potential cancer genes that we have developed by integrating multiple biological evidence, including protein-protein interaction network properties, and sequence and functional features. We detected 55 features that were significantly different between cancer genes and non-cancer genes. Fourteen cancer-associated features were chosen to train the classifier. Four machine learning methods, logistic regression, support vector machines (SVMs), BayesNet and decision tree, were explored in the classifier models to distinguish cancer genes from non-cancer genes. The prediction power of the different models was evaluated by 5-fold cross-validation. The area under the receiver operating characteristic curve for logistic regression, SVM, Baysnet and J48 tree models was 0.834, 0.740, 0.800 and 0.782, respectively. Finally, the logistic regression classifier with multiple biological features was applied to the genes in the Entrez database, and 1976 cancer gene candidates were identified. We found that the integrated prediction model performed much better than the models based on the individual biological evidence, and the network and functional features had stronger powers than the sequence features in predicting cancer genes.

cancer gene, logistic regression, network property, sequence feature, functional annotation

Citation: Liu W, Xie H W. Predicting potential cancer genes by integrating network properties, sequence features and functional annotations. Sci China Life Sci, 2013, 56: 751−757, doi: 10.1007/s11427-013-4500-6

Cancer is an extremely complex genetic disease [1]. The identification of important cancer genes can bring dramatic therapeutic advances and prolong the lives of cancer patients [2]. It has been suggested that while 5%–10% of human genes might contribute to cancer; currently, experimentally validated cancer genes cover only about 1% of the human genome [3]. This fact implies that there are still hundreds or even thousands of cancer genes still to be identified.

Traditional experimental approaches, such as linkage analysis and association studies, are time consuming, laborious and error-prone [4]. Alternatively, with the availability of genome-wide sequences, and genomics and proteomics data, bioinformatics methods have been applied to identify potential cancer genes, significantly reducing the number of candidate genes for further testing [5]. Bioinformatics methods based either on gene annotation and sequence features or on network analysis have provided powerful tools to accelerate cancer gene discovery [6–10]. For example, Furney et al. [6] identified some common structural, functional and evolutionary properties of cancer genes and then used these properties to predict novel cancer genes. Ostlund et al. [7] proposed a network searching method, MaxLink, to find candidate cancer gene based on their connectivity to known cancer genes, and Li et al. [8] integrated network and functional properties to identify cancer genes.

*Corresponding author (email: xhwei65@nudt.edu.cn)

With the accumulation of omics data, it becomes possible to collect, arrange, and integrate diverse biological evidence to build classifiers that can predict novel cancer genes more and more reliably. In this paper, we examined a large number of features from protein sequences, functional annotations and interaction networks and analyzed systematically their roles in the identification of cancer genes. Using four machine learning algorithms, we established a valid classifier that could distinguish cancer genes from other genes, and we evaluated its performance by cross-validation. Finally, the classifier was used to identify potential cancer genes in a gene database.

1 Materials and methods

1.1 Human interaction datasets

Human protein-protein interaction (PPI) datasets were downloaded from the Online Predicted Human Interaction Database (OPID) [11]. We selected the literature-curated interactions from the BIND [12], HPRD [13] and MINT [14] databases. The selected interactions were all obtained experimentally and interactions based only on prediction results were excluded. This interaction dataset was integrated with a previous reported human signaling network that contained 1643 nodes and 5089 signaling regulatory relations [15]. The total number of PPIs and number of unique proteins in the final dataset was 47757 and 10016, respectively.

1.2 Training dataset

A dataset of cancer genes was collected from the Cancer Gene Census [2], the Online Mendelian Inheritance in Man (OMIM) [16], the Network for Cancer Genes (NCG) [17], the COSMIC database and a previously published list of candidate cancer genes [7], as shown in Table 1. The OMIM genes were identified by searching the annotations. By matching the disease annotations of the OMIM genes against cancer-specific terms, the cancer-related genes were identified and added to the set of known cancer genes. The COSMIC database contains the results of large-scale sequencing of tumor samples and provides mutation frequencies for most of the cancer mutated genes to help identify genes that may be critical in the development of human cancers. Genes with 100% mutation frequencies were considered as cancer genes. After the removing redundancies, we obtained an integrated dataset that contained 2104 cancer genes. These genes made up the positive training dataset (Table S1 in Supporting Information).

Table 1 Dataset of cancer genes used in this study

Source	Number
Cancer Gene Census [2]	474
Subset of OMIM [16]	329
NCG [17]	1494
Previous resource [7]	812
COSMIC	218
Total	2104

No verified non-cancer gene dataset was available. Therefore, we constructed a putative non-cancer gene dataset as follows: first, we excluded the Entrez genes [18] that were annotated as essential genes, because it has been reported that these genes have features which differ significantly from both disease-genes and other non-essential genes [19]; next, we removed the disease genes listed in OMIM and called the remaining genes the 'control-gene set'; and finally, we randomly selected genes that were equal in length to the cancer genes in the control-gene set as negative training dataset. The final training dataset consisted of the genes in the sampled negative representative dataset (control group) and the fixed positive cancer dataset (cancer gene group).

1.3 Feature selection

Here, we used a simple and intuitive method, the *F*-score, to measure the power of discrimination of each feature. The *F*-score for feature *i* is defined as

$$F_i = \frac{\left(\overline{x}_i^{(+)} - \overline{x}_i\right)^2 + \left(\overline{x}_i^{(-)} - \overline{x}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left(x_{k,i}^{(+)} - \overline{x}_i^{(+)}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left(x_{k,i}^{(-)} - \overline{x}_i^{(-)}\right)^2}, \quad (1)$$

where $\overline{x}_i$, $\overline{x}_i^{(+)}$, $\overline{x}_i^{(-)}$ denote the mean values of feature *i* in all dataset, positive dataset and negative dataset. $x_{k,i}^{(+)}$ is the value of feature *i* of gene *k* in the positive dataset, and $x_{k,i}^{(-)}$ is the value of feature *i* of gene *k* in the negative dataset. High *F*-scores indicate a discriminative feature.

1.4 Machine learning

Four machine-learning algorithms were investigated: logistic regression, support vector machine (SVM) based on PolyKernel, BayesNet and J48 decision tree, all of which have been widely used for pattern classification and regression problems. The WEKA package [20] was used to build a classifier that could distinguish cancer genes from non-cancer genes, using selected features. A scaling scheme that restricted all entries to be between 0 and 1 was used for every vector, by calculating the (*X*-Min)/(Max-Min) for each feature, where *X* is the feature value, and Min and Max are the minimum and maximum values of *X* in the training dataset.

We can evaluate the performance of the classifiers using 5-fold cross-validation. During the test process, 20% of the genes in the positive and negative datasets were singled out in turn to become the test sample, and the remaining genes

were used as the training set to predict the class of the genes in the test sample. The performance was measured by the analysis of receiver operating characteristic (ROC) curves, which plot the true positive rate against the false positive rate at various thresholds [21]. The area under the ROC curve (AUC) provided the metric for the overall performance of the classifier. The closer the AUC of a test was to 1.0, the higher the overall efficacy of the test.

2 Results and discussion

We have developed an analysis pipeline to identify candidate cancer genes based on PPI networks, sequences and functional features. Our analysis of the differences between cancer genes and non-cancer genes revealed that cancer genes have functional, sequence and network characteristics that are distinctive from those of non-cancer genes. We used the distinctive features to generate a series of valid biological features that could be included in the classifiers. We established the classifiers based on four machine learning methods, namely logistic regression, SVM, BayesNet and decision tree, and evaluated their performance by cross-validation. Finally, we applied the classifier based on the logistic regression method to the Entrez genes and identified a large number of potential cancer genes.

2.1 Extraction of multiple biological features

2.1.1 Network properties

For each node *i* in the PPI network, we defined five measures to assess its topological properties: degree, 1N index, 2N index, shortest distance to cancer genes, and the clustering coefficient. Degree is defined as the number of proteins connected with node *i* and is the most widely used network property. 1N and 2N indexes are defined as the proportion of cancer genes in the neighbors of node *i* and in the neighbors' neighbor of node *i*, as described previously [22]. The shortest distance to cancer genes measure was used to assess the communication efficiency of node *i* to cancer genes in the PPI network. It was assumed that a short distance would correspond to a quick transduction between node *i* and the cancer genes.

A comparison of the five network measures between cancer genes and non-cancer genes is shown in Table 2. In the PPI network, the mean of the degree value for the cancer gene dataset was significantly higher than it was for the non-cancer genes, confirming a previous finding that the degree measure was higher for disease genes compared with non-disease genes [22]. The 1N and 2N indexes for the cancer genes were also significantly higher than those of non-cancer genes, suggesting that the neighbors of a cancer gene are more likely to be cancer genes than non-cancer genes, in agreement with a previous observation about disease genes [22]. We found that, in the control dataset, the shortest path to cancer genes was significantly higher than in the known cancer gene set, indicating that in the PPI network the cancer genes communicated quickly with each other. The clustering coefficient was similar for both the cancer gene and non-cancer gene datasets.

2.1.2 Sequence properties

The sequences of the proteins corresponding to the genes in the cancer and non-cancer datasets were analyzed to extract the main sequence properties. Hydrophobicity was calculated as the sum of hydrophobicity values for the amino acid residues using the Kyte and Doolittle index [23], divided by the number of residues in the corresponding protein sequence. We also computed the frequency of each amino acid in the protein sequence, as the number of each amino acid divided by protein length. The amino acids in each sequence were grouped as tiny, small, aliphatic, aromatic, non-polar, polar, charged or basic [24]. The Pepstats program was used to calculate the protein sequence information statistics (http://emboss.bioinformatics.nl/cgi-bin/emboss/pepstats) for 44 sequence properties, including hydrophobicity, molecular weight, number of residues, pI, amino acid frequencies and other properties.

When the statistics of the sequence properties were compared, we found that the means of 19 sequence features were significantly different between the cancer proteins and non-cancer proteins (P-value<0.05, Table S2). The three properties that differed the most were the number of residues, molecular weight, and the A280 molar extinction coefficient. The cancer proteins tended to be longer than the non-cancer proteins, with mean number of residues of 868 and 559, respectively. This finding is consistent with the description of cancer genes reported previously [5]. Correspondingly, the mean molecular weight of the cancer proteins (96517 Da) was significantly higher than for the non-cancer proteins (66255 Da) (P-value=9.11×10^{-30}). The frequencies of the amino acid in the protein sequences from the two datasets are shown in Figure 1. The differences in the frequencies of the Asn and Leu residues are the most significantly different between the two datasets (P-values= 9.70×10^{-8} and 5.76×10^{-5}, respectively). The cancer proteins tended to have more polar amino acids than the non-cancer proteins, which is similar to the findings of an earlier study

Table 2 Topological features of the cancer gene and the control gene datasets in the PPI network

Network measure[a)]	Cancer	Control	P-value
Degree	18.297	7.774	8.61×10^{-27}
1N index	0.354	0.242	4.85×10^{-26}
2N index	0.343	0.234	3.23×10^{-113}
Shortest distance	1.368	1.804	1.64×10^{-15}
Clustering coefficient	0.124	0.123	0.966

a) All the values are given as means.

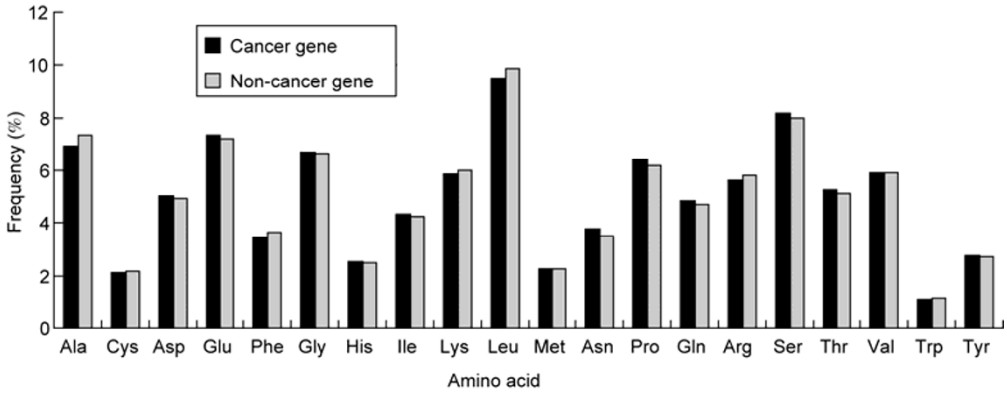

Figure 1 Frequency of amino acids in protein sequences corresponding to the genes in the cancer and non-cancer datasets.

of protein drug targets [24]. In general, cancer proteins have a significantly higher proportion of polar and small amino acids and a lower proportion of non-polar, aliphatic and basic amino acids in their sequences.

2.1.3 Functional properties

We used the Gene Ontology (GO) [25] which provides a controlled vocabulary to describe gene products, under three ontologies, biological process, molecular function, and cellular component, to assign functional properties to the proteins corresponding to the genes in the cancer and non-cancer dataset. In addition, the functional annotations from the Swiss-Prot database, under UP_SEQ_FEATURE and SP_PIR_KEYWORDS, were used to extract cancer-associated functional properties.

Using the DAVID tool [26], we retrieved the GO and Swiss-Prot functional annotation terms for the known cancer genes, and tested them for significance. We also computed the number of genes annotated with these functional terms (the Count value). By selecting the functional categories with P-value <0.001 and Count >300, we identified 32 cancer-associated functional terms (Table S3). We observed strong enrichment (P-value<10^{-10}) for terms such as sequence variant, disease mutation, phosphoprotein and alternative splicing, in good agreement with what is known about cancer-associated processes.

2.2 Computation of *F*-scores for identified features

We detected a total of 55 features that showed significant differences between cancer genes and non-cancer genes, implying their association with cancer. These features include four PPI network properties, 19 sequence features and 32 functional annotation features. The F-scores for all the features were computed to measure their ability to discriminate between the cancer genes and the control genes (see Materials and methods). The 20 features with the highest F-scores are listed in Table 3. The 2N index and 'disease mutation' were the most discriminative features. The cutoff of F-score can be determined according to actual requirement, in order to select the subset of features as the input to build the classifier of cancer genes and non-cancer genes.

2.3 Model establishment and evaluation

Classifiers based on the four methods (logistic regression and support vector machines (SVM), BaysNet and decision tree) were trained with the positive dataset (cancer genes) and the negative dataset (control genes). The PPI network features, sequence properties and functional features were ranked according their F-scores, and a subset of features was selected by setting a cutoff value for the F-scores. The selected features were used as input to the classifiers to distinguish between the cancer and non-cancer genes in the test datasets.

Based on 5-fold cross-validation, we evaluated the performance of the four machine learning classifiers. The AUC for the results of the cross-validation are shown in Figure 2. Not surprisingly, the models with all features as input did not produce the best performances. This might be because of the presence of noisy features that make fitting the hyperplane more complex. Simpler models were obtained by choosing a more relevant subset of features as input and this, in turn, produced a better performance from all four classifiers. Specifically, using an F-score cutoff of 0.03, 14 features were selected to train the classifiers (see the top 14 features in Table 3), including three network properties, two sequence features and nine functional features. The AUC results of the cross-validation of the logistic regression, SVM, Baysnet and J48 tree models were 0.834, 0.740, 0.800 and 0.782, respectively.

The performance of the classifier based on the logistic regression model with a combination of the PPI network, sequence, and functional properties was compared with its performance with each of the property types separately. In all cases, the F-score cutoff was 0.03. The ROCs that we obtained using this model with different input are shown in Figure 3. The corresponding AUC results of the cross-validations for this classifier with all features, and PPI network, sequence and functional features separately were

Table 3 The top 20 features with the highest *F*-scores

Feature type	Feature name	*F*-score
Network	2N index	0.196
Function	Disease mutation	0.170
Function	Sequence variant	0.059
Network	Connectivity	0.045
Sequence	Molecular weight	0.043
Sequence	Residues	0.043
Network	1N index	0.036
Function	ATP-binding	0.034
Function	GO 0005524: ATP binding	0.032
Function	GO 0032559: adenyl ribonucleotide binding	0.032
Function	GO 0001882: nucleoside binding	0.032
Function	GO 0001883: purine nucleoside binding	0.032
Function	GO 0030554: adenyl nucleotide binding	0.032
Function	Phosphoprotein	0.031
Sequence	A280 molar extinction coefficient	0.028
Function	Polymorphism	0.027
Function	GO 0032555: purine ribonucleotide binding	0.025
Function	GO 0032553: ribonucleotide binding	0.025
Function	Nucleotide-binding	0.025
Function	GO 0017076: purine nucleotide binding	0.025

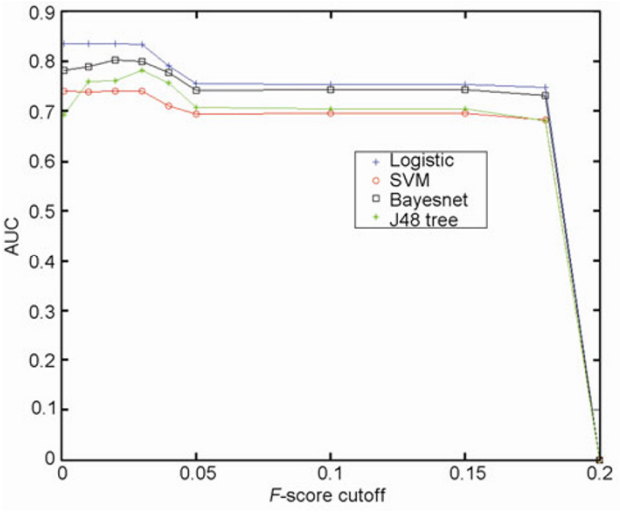

Figure 2 AUC results of the cross-validations for the four classifiers using different *F*-score cutoffs.

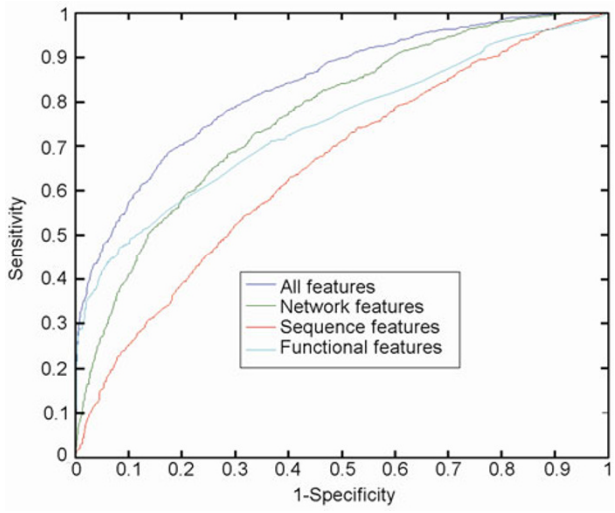

Figure 3 ROCs of the classifier based on logistic regression using multiple biological evidence and individual evidence as inputs.

0.834, 0.768, 0.653 and 0.747, respectively. The results show that the logistic regression model with the combination of features performed better than of the model with the individual feature type. Further, the three feature types differed in their ability to distinguish between the cancer genes and the controls. The sequence features were less sensitive than the network and functional features for the identification of cancer genes.

2.4 Application of the classifier to identify novel cancer genes

We chose the classifier based on the logistic regression method using all the features selected with an *F*-score cutoff of 0.03, and applied it to identify potential cancer genes in the Entrez database [5]. After removing the cancer and non-cancer genes from the gene list, the remaining 6907 genes were imported into the classifier, which predicted that 1976 genes were associated with cancer (Table S4). Some of the predicted cancer genes, which were selected by fixing the following thresholds: (i) cancer linker degree ⩾20; (ii) molecular weight >8000; and (iii) found in disease mutation, are listed in Table 4. These candidate cancer genes will provide a reference dataset for the design of new experiments to improve the understanding of cancer. Further, some of these candidates may be developed into important cancer

Table 4 Cancer gene candidates predicted using the classifier based on the logistic regression method

Gene	Entrez ID	Degree	1N index	2N index	Molecular weight
VCL	7414	130	0.231	0.271	116722
NR3C1	2908	99	0.323	0.338	85659
APP	351	88	0.227	0.310	86943
STAT1	6772	85	0.435	0.375	87334
HD	3064	63	0.270	0.305	347858
ITGB3	3690	54	0.296	0.372	87057
STAT5B	6777	46	0.413	0.424	89865
ITGB2	3689	45	0.244	0.350	84781
PTPRC	5788	44	0.341	0.352	147253
ACTN2	88	44	0.250	0.271	103853
CASK	8573	41	0.122	0.220	104479
JUP	3728	37	0.432	0.354	81744
ITGB4	3691	36	0.417	0.353	202166
VCP	7415	33	0.212	0.265	89321
DNM2	1785	32	0.375	0.331	98064
HSPG2	3339	30	0.167	0.235	468823
COL2A1	1280	29	0.207	0.297	141785
C3	718	28	0.179	0.256	187163
CFTR	1080	28	0.214	0.272	168141
DSP	1832	27	0.222	0.264	260118
IKBKAP	8518	25	0.280	0.311	150253
PKD1	5310	24	0.458	0.323	462415
COL4A1	1282	24	0.250	0.302	160610
PLEC1	5339	23	0.435	0.342	518471
PARD3	56288	23	0.478	0.325	151422
DMD	1756	22	0.182	0.278	425581
RIMS1	22999	20	0.250	0.272	189072
L1CAM	3897	20	0.150	0.292	140002

biomarkers or drug targets.

3 Conclusion

Identification of novel cancer genes is important for understanding the disease mechanism and for the development of cancer therapeutics. In this paper, we integrated multiple types of biological evidence, including sequence properties, PPI network features and functional features to establish prediction models for the identification of potential cancer genes. We confirmed the effectiveness of the prediction model using 5-fold cross-validation. The validation results showed that the integrated prediction models performed much better than the logistic regression model based on the individual biological evidence. Finally, the logistic regression model was applied to the *Entrez* genes to predict candidate cancer genes that could be prioritized for experimental validation.

Because the genes with known functional annotations are the genes that have been most widely investigated, there may be bias in the functional evidence that was used in this study. Therefore, to help eliminate the bias in the individual evidence we integrated three types of biological evidence, including PPI network, functional and sequence features, to identify the potential cancer genes. The method that we proposed here is a powerful tool that can be applied not only to discover unknown cancer genes, but also to provide a comprehensive understanding of cancer from the aspects of sequence, function and PPI networks. In future studies, we will consider combing more data sources, such as transcriptome, proteome, and protein structure data, into the prediction model to further improve its accuracy and application range.

We thank Drs. Zhang JiYang, Wang TengJiao and Xu ChangMing for their excellent advice and assistance as well as all the members in the Bioinformatics Laboratory, College of Mechanical & Electronic Engineering and Automatization, National University of Defense Technology for helpful discussions. This work was supported by the National Natural Science Foundation of China (31000591, 31000587, 31171266).

1 Vogelstein B, Kinzler K W. Cancer genes and the pathways they control. Nat Med, 2004, 10: 789–799
2 Futreal P A, Coin L, Marshall M, et al. A census of human cancer genes. Nat Rev Cancer, 2004, 4: 177–183
3 Strausberg R L, Simpson A J, Wooster R. Sequence-based cancer genomics: progress, lessons and opportunities. Nat Rev Genet, 2003, 4: 409–418
4 Altshuler D, Daly M J, Lander E S. Genetic mapping in human disease. Science, 2008, 322: 881–888
5 Aragues R, Sander C, Oliva B. Predicting cancer involvement of

genes from heterogeneous data. BMC Bioinformatics, 2008, 9: 172
6 Furney S J, Higgins D G, Ouzounis C A, et al. Structural and functional properties of genes involved in human cancer. BMC Genomics, 2006, 7: 3
7 Ostlund G, Lindskog M, Sonnhammer E L. Network-based Identification of novel cancer genes. Mol Cell Proteomics, 2010, 9: 648–655
8 Li L, Zhang K, Lee J, et al. Discovering cancer genes by integrating network and functional properties. BMC Med Genomics, 2009, 2: 61
9 Wang E, Lenferink A, O'Connor-McCourt M. Cancer systems biology: exploring cancer-associated genes on cellular networks. Cell Mol Life Sci, 2007, 64: 1752–1762
10 Milenkovic T, Memisevic V, Ganesan A K, et al. Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data. J R Soc, 2010, 7: 423–437
11 Brown K R, Jurisica I. Online predicted human interaction database. Bioinformatics, 2005, 21: 2076–2082
12 Alfarano C, Andrade C E, Anthony K, et al. The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res, 2005, 33: D418–D424
13 Peri S, Navarro J D, Kristiansen T Z, et al. Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res, 2004, 32: D497–D501
14 Chatr-aryamontri A, Ceol A, Palazzi L M, et al. MINT: the Molecular INTeraction database. Nucleic Acids Res, 2007, 35: D572–D574
15 Cui Q, Ma Y, Jaramillo M, et al. A map of human cancer signaling. Mol Syst Biol, 2007, 3: 152
16 Hamosh A, Scott A F, Amberger J S, et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res, 2005, 33: D514–D517
17 D'Antonio M, Pendino V, Sinha S, et al. Network of Cancer Genes (NCG 3.0): integration and analysis of genetic and network properties of cancer genes. Nucleic Acids Res, 2012, 40: D978–D983
18 Maglott D, Ostell J, Pruitt K D, et al. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res, 2007, 35: D26–D31
19 Tu Z, Wang L, Xu M, et al. Further understanding human disease genes by comparing with housekeeping genes and other genes. BMC Genomics, 2006, 7: 31
20 Frank E, Hall M, Trigg L, et al. Data mining in bioinformatics using Weka. Bioinformatics, 2004, 20: 2479–2481
21 Hanley J A, McNeil B J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 1982, 143: 29–36
22 Xu J, Li Y. Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics, 2006, 22: 2800–2805
23 Kyte J, Doolittle R F. A simple method for displaying the hydropathic character of a protein. J Mol Biol, 1982, 157: 105–132
24 Bakheet T M, Doig A J. Properties and identification of human protein drug targets. Bioinformatics, 2009, 25: 451–457
25 Harris M A, Clark J, Ireland A, et al. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res, 2004, 32: D258–D261
26 Huang da W, Sherman B T, Lempicki R A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res, 2009, 37: 1–13

Supporting Information

Table S1 List of cancer genes in the positive training dataset

Table S2 List of sequence features that were significantly different in the cancer and non-cancer proteins

Table S3 Functional annotation terms that were significantly enriched in cancer genes

Table S4 List of potential cancer genes identified by the classifier based on the logistic regression method using all the features (*F*-score cutoff was 0.03)

The supporting information is available online at life.scichina.com and www.springerlink.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.